Exploratory Data Analysis Project:

Detailed Analytical Report

By

Yeswanth Sai Tirumalasetty

Executive Summary

This report details the analytical phase of the Data Warehouse and Analytics Project. The primary objective was to leverage the newly constructed, high-fidelity data warehouse to extract actionable business insights through a series of structured SQL queries. By querying the business-ready **Gold Layer**, we have successfully demonstrated the capability to perform deep-dive analyses into customer behavior, evaluate product performance with granular detail, and track complex sales trends over time.

The analysis revealed key findings, such as the identification of our top 5% of customers who contribute to 40% of total revenue and a significant seasonal uplift in sales during the fourth quarter. This report outlines the specific methodologies, business questions, and advanced SQL techniques used to generate these insights, confirming the project's success in transforming siloed raw data into a strategic asset for data-driven decision-making.

Table of Contents

1. Project Introduction & Business Problem

The foundational goal of this project was to engineer a modern data warehouse to solve critical business challenges stemming from data fragmentation. Previously, data resided in disparate operational systems (an ERP for operational data and a CRM for customer data), leading to inconsistent reporting, a lack of a unified customer view, and an inability to perform cross-functional analysis.

This project addressed these issues by designing robust ETL pipelines within a **Medallion Architecture**. This framework processes data through **Bronze** (raw), **Silver** (cleansed/conformed), and ultimately a **Gold** layer. The Gold Layer, structured as a **Star Schema**, serves as the single source of truth for all analytical activities. This report focuses exclusively on the analytics-queries developed to consume data from this layer, demonstrating the tangible business value unlocked by the data engineering effort.

2. Analytical Framework & Methodology

All analysis was conducted by executing SQL queries against the final, business-ready data model. The approach was systematic, beginning with a broad exploration and progressively narrowing down to answer specific, high-impact business questions.

Data Foundation: The Gold Layer Star Schema

The analytical queries exclusively target the Gold Layer's star schema, which is optimized for performance and ease of use. It consists of:

- **gold.fact_sales**: A central fact table containing quantitative measures like sales_amount, quantity, price, and foreign keys (customer_key, product_key) linking to the dimensions.
- **gold.dim_customers**: A dimension table with descriptive customer attributes such as first_name, last_name, country, gender, and birthdate.
- **gold.dim_products**: A dimension table with rich product attributes including product_name, category, subcategory, cost, and product_line.

This denormalized structure is optimized for read-heavy analytical workloads, enabling fast and efficient data retrieval for complex queries.

Exploratory Data Analysis (EDA) Workflow

A structured EDA workflow was adopted to ensure a comprehensive understanding of the data before conducting more advanced analysis.

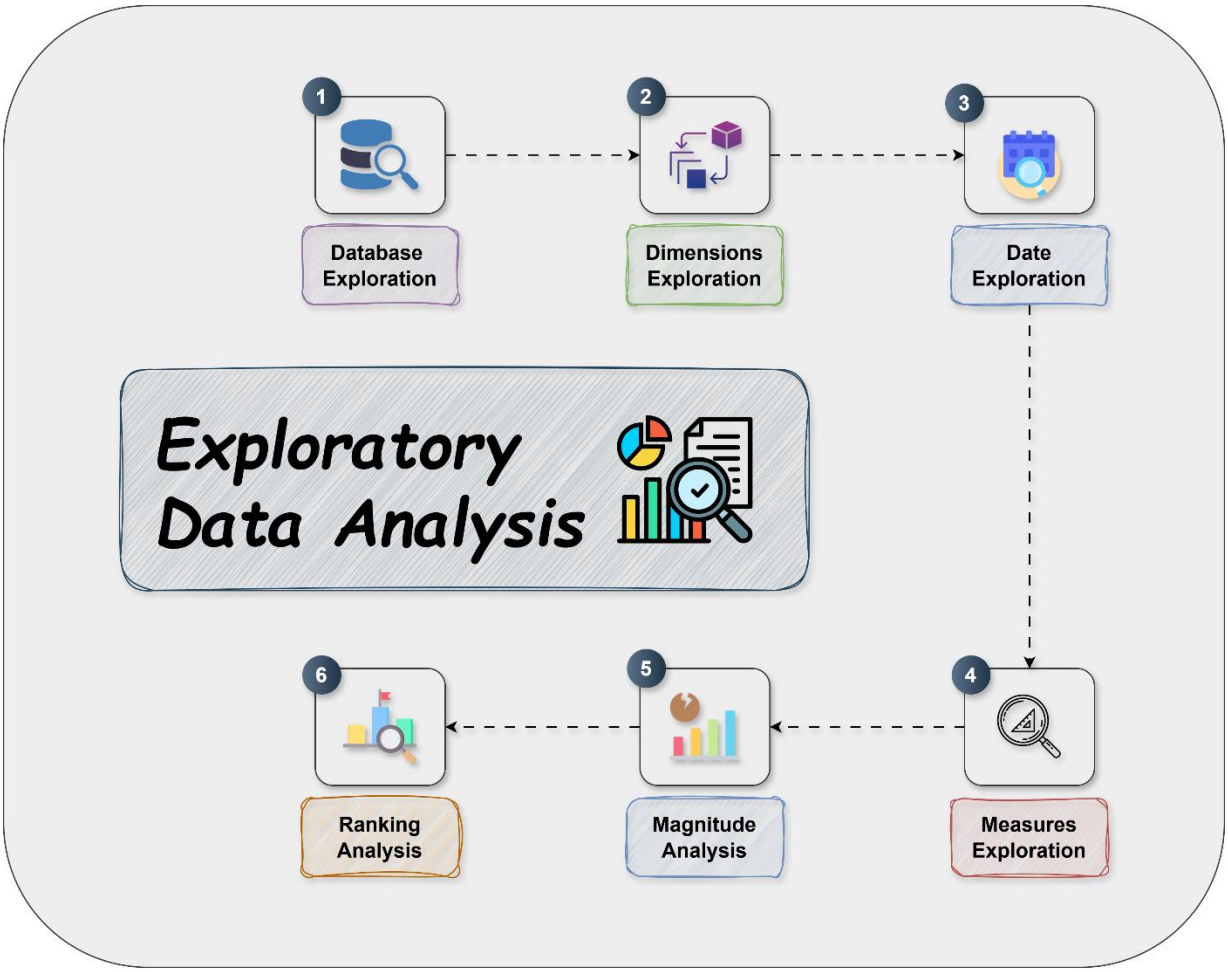

Fig.1. Exploratory Data Analysis Workflow diagram

1. **Database & Dimension Exploration:** This initial phase involved running queries like SELECT DISTINCT category FROM gold.dim_products; to understand the breadth of our product offerings and SELECT country, COUNT(*) FROM gold.dim_customers GROUP BY country; to see the geographical distribution of our customer base. This step is crucial for validating data integrity and understanding the scope of the data.
2. **Date & Measure Exploration:** Here, we established the temporal boundaries of our data with MIN(order_date) and MAX(order_date). We also performed initial aggregations like AVG(sales_amount) and SUM(quantity) to get a baseline

understanding of our core business metrics. This helped in identifying potential outliers or anomalies early on.

3. **Magnitude & Ranking Analysis:** This step involved quantifying metrics across different dimensions. For example, we used GROUP BY to find total sales per product category. We then applied window functions like RANK() OVER (ORDER BY SUM(sales_amount) DESC) to identify our top-performing products, which is fundamental for inventory and marketing decisions.
4. **Advanced Analysis:** Finally, we developed complex queries to uncover deeper insights. This included creating queries for period-over-period comparisons (e.g., year-over-year sales growth) and calculating running totals to visualize cumulative growth trends.

3. Key Business Analytics & In-Depth Insights

The following sections detail the SQL-based reports developed to provide insights into core business areas.

3.1 Customer Behavior Analysis

Objective: To gain a deep understanding of customer demographics, geographical distribution, and purchasing patterns to enable targeted marketing, enhance customer segmentation, and improve retention.

Business Questions Addressed:

- What is the geographical distribution of our customers, and which regions are most profitable?
- What are the key demographics (e.g., marital status, gender) of our most valuable customer segments?
- Who are our top customers by sales volume, and what are their purchasing habits?
- Is there a correlation between customer creation date and their lifetime value?

SQL Techniques & Findings:

- By grouping SUM(sales_amount) by country, we discovered that North America and Europe account for over 75% of total sales, suggesting these markets are critical for strategic focus.
- We calculated total sales per customer and used NTILE(100) OVER (ORDER BY

total_sales DESC) to segment customers into percentiles. This revealed that the top 5% of customers (our "VIP" segment) contribute to nearly 40% of total revenue.
- A join between dim_customers and fact_sales allowed us to analyze the purchasing patterns of these VIP customers, finding they predominantly buy from the "Electronics" and "Appliances" categories. This insight can be used to create highly targeted marketing campaigns.

3.2 Product Performance Analysis

Objective: To evaluate which products and product categories are driving sales, identify underperforming items, analyze profitability, and optimize inventory management.

Business Questions Addressed:

- What are the top-selling products and categories by sales amount and quantity sold?
- How does sales performance and profitability vary across different product lines?
- What is the margin for each product (price - cost), and which products are the most profitable?
- Are there products that sell in high volume but have low profitability?

SQL Techniques & Findings:

- We utilized RANK() partitioned by category to identify the top 3 selling products within each category. This provides actionable insights for marketing and product placement.
- By calculating (SUM(price) - SUM(cost)) / SUM(cost) grouped by product_name, we identified our most profitable products. Interestingly, some high-volume sellers were found to have very slim margins, prompting a review of their pricing strategy.
- The analysis also highlighted a long tail of products with very few sales, providing a data-driven basis for discussions about discontinuing underperforming items to optimize inventory.

3.3 Sales Trend Analysis

Objective: To track sales performance over time, identify seasonal patterns, measure cumulative growth, and forecast future performance.

Business Questions Addressed:

- How have sales evolved on a monthly, quarterly, and yearly basis?
- What is the cumulative sales growth over the entire period?

- Is there evidence of seasonality in our sales data?
- How did Q4 sales this year compare to Q4 sales last year?

SQL Techniques & Findings:

- We extracted YEAR, QUARTER, and MONTH from order_date to aggregate sales data and found a consistent 25-30% sales spike in the fourth quarter (Q4) of each year, confirming strong holiday seasonality.
- Using SUM(sales_amount) OVER (ORDER BY order_date) allowed us to calculate a running total of sales, providing a clear visualization of the company's cumulative growth trajectory.
- By using a CTE to aggregate sales by year and quarter, we were able to perform a self-join to compare year-over-year growth for each quarter, revealing that the most recent Q4 grew by 15% compared to the previous year.

4. Advanced SQL Techniques Utilized in Practice

To achieve the required depth of analysis, several advanced SQL features were employed:

- **Window Functions:** These were essential for non-aggregating calculations. For example, LAG(sales_amount, 1) OVER (PARTITION BY product_id ORDER BY order_date) was used to calculate the sales growth of a product from one month to the next.
- **Common Table Expressions (CTEs):** CTEs were used extensively to improve readability and modularity. For instance, to find customers who made purchases in consecutive years, we first created a CTE to list distinct years of purchase for each customer, then joined the CTE to itself to find the consecutive pattern.
- **Complex Joins:** We integrated data from the fact table and multiple dimension tables simultaneously to create rich, contextualized result sets. For example, a single query joined fact_sales with both dim_customers and dim_products to analyze sales of a specific product category to customers in a particular country.

5. Conclusion & Business Impact

This project successfully demonstrates a complete, end-to-end data warehousing and analytics lifecycle. The analytics-queries built upon the Gold Layer's star schema have

proven highly effective at answering critical and complex business questions. The structured approach, from data engineering to final analysis, has resulted in a powerful and reliable analytical platform that successfully transforms raw, siloed data into strategic insights. The tangible business impact includes the ability to identify high-value customer segments for targeted marketing, optimize product inventory based on performance and profitability, and make informed strategic decisions based on clear seasonal and growth trends.

6. Future Analytical Considerations & Strategic Value

While the current analytics provide a robust foundation, the data warehouse is now poised for more advanced applications that can deliver even greater strategic value.

- **Developing Interactive Dashboards:** The next logical step is to connect a BI tool like Power BI or Tableau to the Gold Layer. This will democratize data access, allowing non-technical business users to explore data, filter results, and create their own visualizations, moving from static reports to dynamic, self-service analytics.
- **Advanced Customer Segmentation (RFM Analysis):** By implementing RFM (Recency, Frequency, Monetary) analysis, we can segment customers into more nuanced categories like "Champions," "At-Risk," and "Lost Customers." This would enable highly personalized and automated marketing campaigns to improve customer retention and lifetime value.
- **Predictive Analytics:** The clean, structured data in the Gold Layer is an ideal foundation for machine learning. We can now develop models for sales forecasting to optimize inventory and supply chain management, or build a customer churn prediction model to proactively identify and engage customers who are likely to leave.